SENSITIVITY ANALYSIS IN STATISTICS

Ali S. Hadi and Hans Nyquist

94 – 09

WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 94-09
Statistics and Econometrics series 06
April 1994

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe, Spain
Fax: (341) 624-9849

SENSITIVITY ANALYSIS IN STATISTICS

Ali S. Hadi and Hans Nyquist[1]

Abstract:

Statistical models are simplification of reality; we rarely expect the model to be exactly true. Nevertheless, when we select a statistical technique and a perform statistical inference, we often act as if the model is true. This is often justified by claiming that "small" deviations from the model cause only "small" deviations from the theoretical properties of the selected inferential techniques or cause only minor changes in the results produced by the inference. Unfortunately, this argument need not be true. In some applications apparently small changes in a model, a model assumption, or a data point, can have very large effects on the results. For this reason, statistical analysis is viewed in this paper as a cyclical process. Such a process takes inputs and produces outputs in an iterative or cyclical way; a way in which the outputs can be used to diagnose, validate, criticise, and possibly alter the inputs. We also describe a general framework, referred to as the sensitivity function, for assessing the sensitivity of the outputs to small changes in the input at a given cycle of the statistical process. We give several examples from various areas in statistics illustrating the general applicability of the sensitivity function and show how and where the sensitivity function fits into the statistical cycle. Some applications of the sensitivity function lead to known statistical techniques, while other applications produce new ones.

Keywords:

Binomial distribution, Frechet distance, influence function, regression diagnostics, sensitivity function, Score test, Statistical cycle, Wald test, Likelihood ratio test.

[1] Ali S. Hadi, Department of Statistics, Cornell University, 358 Ives Hall, Ithaca, N.Y. 14853-3901, USA (E-mail: ali-hadi@cornell.edu). Hans Nyquist, Department of Biometry and Forest Management, Swedish University of Agricultural Sciences, S-901 83 Umeå, Sweden (E-mail: hans.nyquist@biom.slu.se.bitnet). Part of this work has been done while Ali S. Hadi was visiting Departamento de Estadística y Econometría, Universidad Carlos III de Madrid.

1. Introduction

A statistical analysis is viewed here as a process which takes inputs and produces outputs in an iterative and cyclical way; a way in which the outputs can be used to diagnose, validate, criticise, and possibly alter the inputs. Figure 1, which is adapted from Box (1979, 1980), illustrates this cyclical process. The typical inputs in a statistical process include: (a) the subject matter theories or hypotheses, (b) the chosen model(s), (c) the data (obtained, e.g., from a survey or from a designed experiment), and (d) the selected statistical technique(s). The selected statistical techniques also require certain assumptions which we refer to as auxiliary assumptions, since they are made only for the convenience of the statistician. Typical outputs of the statistical process include estimated parameters, confidence regions, test statistics, etc.

The objectives of this paper are: (a) to describe the statistical process and emphasize its iterative nature (Section 2), (b) to describe a general framework, referred to as the sensitivity function, for assessing the sensitivity of the outputs to small changes in the input of a given statistical process (Section 3), and (c) to give several examples from various areas in statistics illustrating the general applicability of the sensitivity function and show how and where the sensitivity function fits into the statistical cycle (Section 4). We shall see that some applications of the sensitivity function lead to known statistical techniques, yet other applications produce new ones. A summary and concluding remarks are given in Section 6.

2. The Iterative Nature of the Statistical Process

A more detailed description of the statistical cycle of Figure 1 is given in Figure 2. Typically, we start a statistical analysis with a population consisting of N (possibly unknown and/or infinite) elements. Ideally, the population ought to be well-defined, but well-defined populations are sometimes hard to come by. To obtain information about some unknown characteristics (e.g., parameters) of the population, we collect a sample of size n elements. The sample size may or may not be determined in advance. Also, a sampling procedure or design has to be chosen. Ideally, the sampling procedure should ensure: (a) that the sample be a representative of the population, (b) that all sample elements be drawn from the same population, and (c) that the sample elements be independently drawn. The last two conditions, which are referred to as independently and identically distributed samples, are required by most of the commonly used statistical techniques, chief among them is the maximum likelihood method.

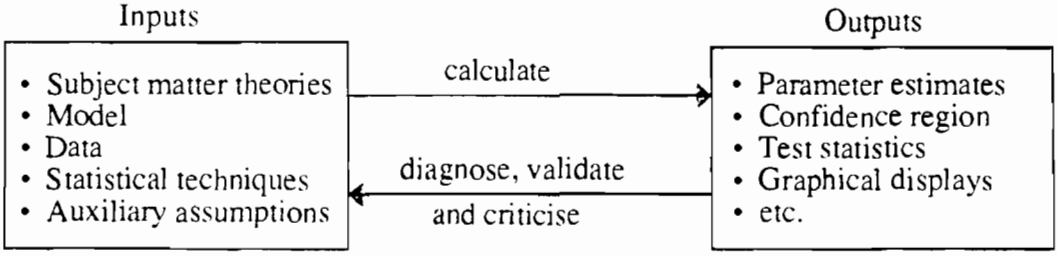

Figure 1. A schematic illustration of the statistical cycle.

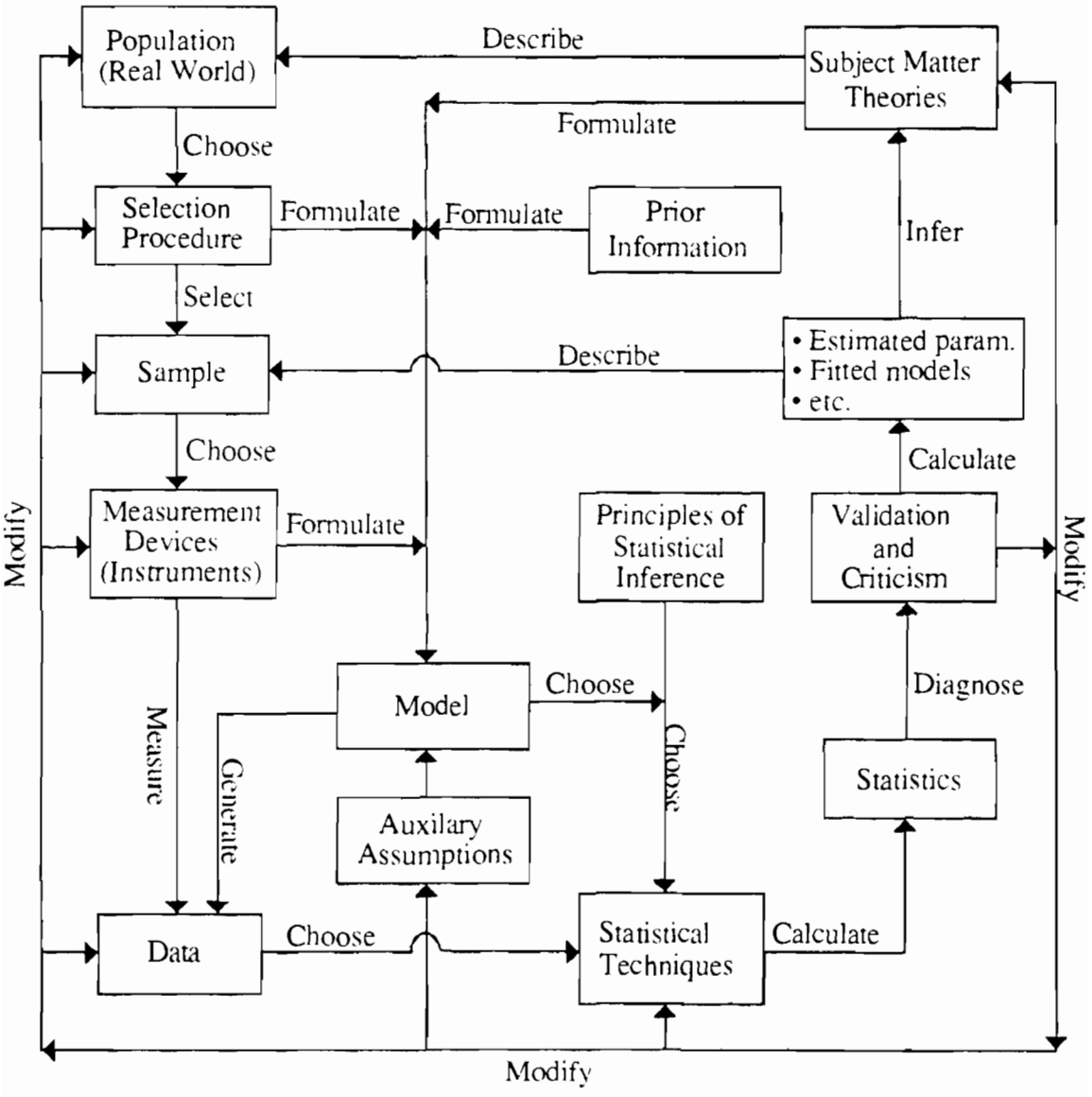

Figure 2. A detailed chart illustrating the iterative statistical process.

Particular characteristics of the sample are then measured using some instruments or measurement devices (e.g., questionaires, balances, laser, etc.). The resultant measurements constitute the data. Aside from possibly having some prior information about the population and/or the parameters, all the empirical information we have about the population are contained in the data. In that sense, the data are the most important component of a statistical analysis.

Most, if not all, statistical techniques are based on the premise that all data points (observations) play an equal role in determining the results. Unfortunately, this is seldom the case in practice. In some applications one or few data points can have a substantial influence on determining the results of an analysis. An extensive discussion concerning influential data points in linear regression has been in existence for a long time, see for example, Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1982), Atkinson (1985), and Chatterjee and Hadi (1988). In addition to these books, a large amount of research papers have been published. On the other hand, few results about influential data points in fields other than linear regression analysis have been published. Pregibon (1981) and Chatterjee and Hadi (1988, Chapter 8) treat the analysis of generalized linear models. An excellent review of diagnostics for the generalized linear models and extensions to more general models are given by Davison and Tsai (1992).

A model can be thought of as a description of the mechanism that generated the data. Ideally, the model is formulated based on:

(a) our knowledge of the subject matter theories or hypotheses,

(b) the availability of prior information, for example, if a prior distribution for the parameters is given, Bayesian models may be chosen,

(c) known characteristics of the sample selection procedure, for example, one sample selection procedure may produce independent elements while another may produce dependent elements (e.g., neighboring pixels in a satellite image are supposed to be dependent),

(d) the way the measurement were taken., for example, a person's height may be measured by a tape (yielding a conceptually continuous data) or by simply classifying the person into one out of predetermined height classes. These two methods of measuring height may lead to different models (e.g., normal for the former and probit for the latter).

The model we use is, however, in many respects a simplified representation of reality; we rarely expect the model to be exactly true. Nevertheless, when selecting a technique for statistical inference, and when performing the inference, we often act as if the model is true. This is often justified by claiming that "small" deviations from the model cause only "small" deviations from the theoretical properties of the selected inferential techniques or cause only

minor changes in the results produced by the inference engine. Unfortunately, this argument need not be true. In some applications an apparently small deviation from the model can have very large effects on the results.

We next choose the statistical technique that is appropriate for the analysis. This choice, which may depend on the type of data and/or the form of the model, is guided by the principles of statistical inference (e.g., unbiasdness, minimum variance, consistency, efficiency, sufficiency, etc.).

Most, if not all, statistical techniques require some auxiliary assumptions (e.g., normality, independence, constant-variance, etc.). The effects of a misspecified model or an invalid assumption are again rather well known in the case of linear regression (many results appear in standard text books) but only few results are published for other models. An important exception here is the field of robust statistics (see e.g. Huber, 1981 and Hampel et al., 1986) which can be said to have emerged from the problem of erroneous distributional assumptions.

Thus, the subject matter theories, the prior information, the formulated model, the measured data, the selected statistical techniques, and the associated assumptions constitute the initial input to the statistical process.

It can easily be seen that countless number of errors can creep into the statistical analysis at various stages of the process. The following is by no means an exhaustive list of such possible errors:

- The population was vaguely defined and, as a result, some elements of the sample (possibly unknown in number and in identity) were drawn from populations different from the target population. This type of error is usually referred to as a contamination error.
- Numerous errors can occur as a result of a badly designed experiment or a questionaire (e.g., too many or too vague questions, questions leading to response and/or non-response biases, incorrect answer due to interviewer bias, the experiment was designed based on the wrong model, etc.)
- The instruments by which we measure the characteristics of the elements are imprecise causing what are referred to as measurement errors.
- Some data values were incorrectly coded/decoded at the source and/or incorrectly entered into the computer (e.g., misplacing a decimal point or transposing two digits).
- Because our knowledge of the subject matter is limited, vague, or inaccurate, an incorrect model can initially be chosen.
- Some of the auxiliary assumptions do not hold (e.g., the data are not independent, not normally distributed, not symmetric, not continuous, etc.).

- The statistical technique is chosen based on optimal statistical properties. These properties are usually contingent on the correct choice of the model and on the validity of the assumptions. These optimal properties may be lost if an incorrect model or an incorrect assumption has been originally chosen.

Obviously, some of these errors are inherent characteristics of the statistical process. In that sense, statistical process is a stochastic process. Nevertheless, care has to be taken so as to minimize the number and magnitude of these errors throughout the statistical process.

We can now enter the data into the computer (hopefully without errors) and choose our favorite statistical package or write our own programs to calculate some preliminary results. The results can take many forms such as numerical summaries, tables, charts, graphs, etc. Before one could make any conclusions about the characteristics of the population under study, one must first use the preliminary results to validate, criticise, and diagnose problems with the various inputs of the process. Here where sensitivity analysis plays its major role. At this stage in the process, one could check the sensitivity of the obtained results to small changes in the inputs. Several "what-if?" type of questions could be asked. For example:

- Given the uncertainty about the form of the model, are the results sensitive to small changes in model specifications?
- Given the uncertainty surrounding the model, do we obtain different results if we try different statistical techniques (e.g., using maximum likelihood or least absolute deviations instead of least squares)? For example, it would not be clear how to estimate the center of the distribution of a univariate data when the distributional form is unknown and where the mean (the least squares estimate) and the median (the least absolute deviations estimate) are substantially different from each other.
- Are the assumptions required by the statistical technique valid? What are the effects on the results if some of these assumptions are invalid?
- Does the data set appear to be homogeneous or does it contain unexpected clusters or exhibit unexpected patterns and structure?
- Do the data contain outliers and/or influential observations? If so, these observations must be examined thoroughly before a decision can be made as to what to do with them.
- Does the proposed model adequately fit the data? If not, what can be done about it? For example, should we transform the data to conform with the model or should we search the subject matter knowledge for an alternative form of the model?

This motivates the introduction of various methods for assessing the sensitivity of the results caused by a questionable model, a questionable assumption, or a questionable data point. These aspects, however, have been often treated separately in the statistical literature. Often the sensitivity of parameter estimates are considered, but also values of other statistics appear, e.g. estimators variances and various goodness-of-fit statistics. Furthermore, the effects are considered either asymptotically or in finite samples. For each combination of statistic, aspect under consideration, and sample size, several measures for assessing sensitivity can be defined. This gives a multitude of possibilities for analyzing effects of a questionable model, a questionable assumption, or a questionable data point. Many of the proposed measures show, however, a similar structure. One purpose of this paper is to describe a structure to which many sensitivity measures apply. This structure is presented in Section 3 and applied to several areas of statistics in Section 4.

The results of the sensitivity analysis can then be used to modify any or all of the inputs, for example:

- an auxiliary assumption may be relaxed or replaced,
- the subject matter theories, hence the model, may be modified,
- a modification of the model or a relaxation of an assumption may lead to the selection of a different statistical technique,
- erroneous data points can be corrected, the data may be transformed, outliers may be down-weighted or discarded, etc.,
- the sample size may be too small, the sample may prove to be inadequate or not representative of the population, the questionaire needs to be redesigned, etc., and
- the population may have to be redefined (e.g., it may be easier to deal with two homogeneous sub-populations rather than dealing with one population containing two distinctive groups).

Several iterations may be needed before one arrives at results which are insensitive to all questionable input items. Then, and only then, one should compute various statistics of interest. These statistics can be used to describe the sample and to obtain the final results (e.g., estimated parameters, fitted model, various test statistics, etc.). These results can then be used for making appropriate inferences about the characteristics of the population under study.

3. The Sensitivity Function

Suppose that the model under study, M_0, is embedded into a larger class of models, M, so that M itself is a parametric model. Let ϕ be a parameter that indexes the members of M

such that M_0 is obtained for $\phi = 0$ and M_ϕ is an arbitrary member of M. Suppose further that $T(M_\phi)$ is a statistic under consideration and $P_T(M_0)$ is a specified property of T under the model M_0. (For simplicity of notation, we write $T(M_\phi)$ as T). Examples of the properties of T that may be considered include: ,

- an observed value of T,
- the population value of T,
- the expected value of T,
- the variance of T,
- the sampling distribution of T,
- the asymptotic distribution of T,
- an empirical distribution of T, and
- the likelihood evaluated at an observed value of T,

The sensitivity of a specific property of T with respect to changes in ϕ is obtained by comparing $P_T(M_\phi)$ and $P_T(M_0)$. In the simplest cases where ϕ is a scalar and $P_T(M_\phi)$ is finite dimensional, the difference $\rho(\phi) = P_T(M_\phi) - P_T(M_0)$ is usually considered. More generally, the comparison is made in terms of $\rho(\phi) = d(P_T(M_\phi), P_T(M_0))$, where $d(\cdot, \cdot)$ is some suitable function. If, for example, $P_T(M_\phi)$ is a distribution function, it is convenient to let $d(\cdot, \cdot)$ be a metric defined on a space of probability distributions (examples include the Frechet, Levy, and Prohorov distances); or if $P_T(M_\phi)$ is an observed vector-valued function of T, a suitable semi-norm of $(P_T(M_\phi) - P_T(M_0))$ may be used. Some examples of $d(\cdot, \cdot)$ are given in Section 4.

With this definition, $\rho(\phi)$ is a measure of the change in P_T when ϕ is changed. In order to describe the behaviour of $\rho(\phi)$ in a neighborhood of $\phi = 0$, one may take a Taylor series expansion of $\rho(\phi)$ around $\phi = 0$, namely,

$$\rho(\phi) = \rho(0) + \phi\,\rho'(0) + \phi^2\rho''(0)/2 + \ldots, \tag{2.1}$$

where the constant term $\rho(0)$ is the value of $\rho(\phi)$ at M_0, and $\rho'(0)$ and $\rho''(0)$ are the first and second derivatives of $\rho(\phi)$ with respect to ϕ evaluated at $\phi = 0$, respectively. Each term in the expansion contribute to the behaviour of $\rho(\phi)$ in the neighborhood of $\phi = 0$, and hence to a description of the sensitivity of P_T to deviations in ϕ from $\phi = 0$. We now focus our attention on the first order term, called the sensitivity function in Nyquist (1992), which is defined by

$$SF(M, T, P) = \lim_{\phi \to 0} d\{P_T(M_\phi), P_T(M_0)\} \,/\, \phi, \tag{2.2}$$

provided that the limit exists. Thus, the sensitivity function can be interpreted as the relative change in $\rho(\cdot)$ under a small change in ϕ, i. e., it measures the local sensitivity of P_T. In many applications, terms of higher order can be neglected, but this has to be checked in each separate case. Provided that higher order terms in the expansion (2.1) can be neglected, large values of the sensitivity function indicate a large local sensitivity to changes in ϕ, while values close to zero indicate only a small local sensitivity.

4. Some Applications of the Sensitivity Function

In this section we give examples illustrating the applications of the sensitivity function in various areas of statistics.

Example 4.1. The Influence Function. This example shows that the influence function (Hampel, 1968, 1974) is a special case of the sensitivity function. Suppose that P_T is the population value of a finite dimensional statistic T and that the model M_ϕ, is represented by the cumulative distribution function, $M_\phi = (1 - \phi)F + \phi\,\delta_z$, where F is the distribution function for the model M_0 and δ_z is the distribution that assigns point mass 1 at the point z in the sample space. Representing the statistics as functionals, we have $P_T(M_\phi) = T((1 - \phi)F + \phi\,\delta_z)$. With $\rho(\phi) = P_T(M_\phi) - P_T(M_0)$, the sensitivity function reduces to

$$\begin{aligned} SF(M, T, P) &= \lim_{\varphi \to 0} \{T((1 - \phi)F + \phi\,\delta_z) - T(F)\} / \phi \\ &= IF(F, T, z), \end{aligned}$$

where $IF(F, T, z)$ is the influence function, a tool which is extensively used in robust statistics (see, e. g., Huber (1981) and Hampel et al. (1986)).

Example 4.2. Finite Sample Approximations of the Influence Function. We consider the same settings as in the Example 4.1, except that here we take $P_T(M_\phi) = t(\phi)$ to be the observed value of a finite dimensional statistic T with a weight $1 - \phi$ assigned to the ith observation and all other observations are assigned a unit weight. This means, e.g., that $t(0)$ and $t(1)$ are the observed values of T for the complete data and for the reduced data (when the ith observation has been removed), respectively. Selecting $\rho(\phi) = t(\phi) - t(0)$, the application of the sensitivity function approach allows assessing the effects of infinitesimal perturbations of the ith data point. Provided that the derivative exists, we obtain $SF(M, T, t) = t'(0)$. In particular, $n\,t'(0)$ is the so called empirical influence function and $-(n - 1)t'(\phi)$ for some $\phi \in (0, 1)$, equals the sample influence function,

$(n-1)(t(0)-t(1))$. Hence, $t'(0)$ describes local changes in t at the fitted model while $t'(1)$ describes local changes in t after the ith observation has been removed. Thus, this example shows that the finite sample approximations of the influence function are special cases of the sensitivity function.

Example 4.3. Elliptically Symmetric Distributions. Let $\mathbf{X}$ be an $n \times k$ matrix whose rows are drawn independently from an elliptically symmetric distribution F (e. g., multivariate normal distribution) with center μ and finite scale Σ. Let $\mathbf{x}_i$ be the transpose of the ith row of $\mathbf{X}$. One statistic of interest here is the maximum likelihood estimate (MLE) of μ which is given by the sample mean $\mathbf{T} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$.

Suppose now that the distribution is perturbed by adding a distribution with mass at the point $\mathbf{z}$ and that the weight ϕ is assigned to that distribution. Of interest is the assessment of the influence of $\mathbf{z}$ on the asymptotic distribution of $n^{-1/2}\mathbf{T}$. Setting $P_{\mathbf{T}}(M_\phi) = F(\phi)$, the influence of $\mathbf{z}$ on this distribution can then be measured by $d\{F(\phi), F(0)\}$ for some appropriate distance d defined on a space of probability distributions. The associated sensitivity function can be written as

$$SF(M, \mathbf{T}, \mathbf{P}) = \lim_{\phi \to 0} \frac{1}{\phi} d(F(\phi), F(0)). \tag{4.1}$$

As an example, let us take $d(\cdot, \cdot)$ to be the Frechet distance (Frechet 1957).

For two random variables W and V with distribution functions G and H, the Frechet distance between G and H is defined by

$$d(G, H) = \{ \min_{W, V} E \| W - V \|^2 \}^{1/2}, \tag{4.2}$$

where the minimization is taken over all random variables W and V having distributions G and H, respectively. Dowson and Landau (1982) show that, if G and H are elliptically symmetric, (4.2) can be written as

$$d(G, H) = \{ \| \mu_W - \mu_V \|^2 + \mathrm{tr}[\Sigma_W + \Sigma_V - 2 (\Sigma_W \Sigma_V)^{1/2}] \}^{1/2}, \tag{4.3}$$

where μ_W, μ_V, Σ_W, and Σ_V are the means and covariance matrices of the random variables W and V, respectively. The square-root is taken to be the positive square-root.. When G and H in (4.3) are univariate, the Frechet distance takes the simple form

$$d(G, H) = \{ (\mu_W - \mu_V)^2 + (\sigma_W - \sigma_V)^2 \}^{1/2}. \tag{4.4}$$

The Frechet distance between the two elliptically symmetric distributions lends itself to nice interpretations. The first term on the right-hand-side of (4.3) defines a metric on the space of all mean vectors of order $k \times 1$ and the second term defines a metric on the space of all covariance matrices of order $k \times k$.

Now, in our case G and H are the asymptotic distributions of $n^{-1/2}\mathbf{T}$ when the sample is drawn from the perturbed and the unperturbed elliptically symmetric distributions, respectively. Thus, $G = F(\phi)$ is the normal distribution with mean vector $\mu(\phi) = (1 - \phi)\mu + \phi\mathbf{z}$ and covariance matrix $\Sigma(\phi) = (1 - \phi)\Sigma + \phi(\mathbf{z} - \mu)(\mathbf{z} - \mu)^T$, and $H = F(0)$ is the normal distribution with mean vector $\mu(0) = \mu$ and covariance matrix $\Sigma(0) = \Sigma$. Therefore, (4.3) becomes

$$\begin{aligned} d(F(\phi), F(0)) = \Big\{ & \| (1 - \phi)\mu + \phi\mathbf{z} - \mu \|^2 \\ & + \operatorname{tr}[(1 - \phi)\Sigma + \phi(\mathbf{z} - \mu)(\mathbf{z} - \mu)^T + \Sigma] \\ & - 2 \operatorname{tr}[(((1 - \phi)\Sigma + \phi(\mathbf{z} - \mu)(\mathbf{z} - \mu)^T)\Sigma)^{1/2}] \Big\}^{1/2}. \end{aligned} \tag{4.5}$$

After some algebra, the first term in (4.5) can be written as $\phi^2(\mathbf{z} - \mu)^T(\mathbf{z} - \mu)$ and the last two terms as

$$\operatorname{tr}\left\{\left[(\mathbf{z} - \mu)(\mathbf{z} - \mu)^T - \Sigma\right]\left[\left((1 - \phi)\Sigma + \phi(\mathbf{z} - \mu)(\mathbf{z} - \mu)^T\right)^{1/2} + \Sigma^{1/2}\right]^{-1}\right\}^2.$$

Substituting these in (4.1) and taking the limit, we obtain

$$SF(M, \mathbf{T}, \mathbf{P}) = \left[\mathbf{u}^T\mathbf{u} + \frac{1}{4}\operatorname{tr}\left\{\left(\mathbf{u}\mathbf{u}^T - \Lambda\right)^2\Lambda^{-1}\right\}\right]^{1/2}, \tag{4.6}$$

where Λ and Γ are the matrices of eigenvalues and associated eigenvectors of Σ, respectively, and $\mathbf{u} = \Gamma^T(\mathbf{z} - \mu)$. In the univariate case, (4.6) reduces to

$$SF(M, \mathbf{T}, \mathbf{P}) = \left[(z - \mu)^2 + \left((z - \mu)^2 - \sigma^2\right)^2 / 4\sigma^2\right]^{1/2}.$$

Equation (4.6) combines two distinct measures of the effects of the observation $\mathbf{z}$: the effect on the mean vector and the effect on the covariance matrix. It shows that these effects would be severe if (a) $\mathbf{z}$ is far from μ and (b) $\mathbf{z}$ lies in the direction of any eigenvector associated with a small eigenvalue of Σ. In other words, if the perturbation occurs far from the mean in a direction where the data are least variable.

A finite sample version of this measure is obtained when μ and Σ are replaced by appropriate (possibly robust) estimates (e.g., the minimum volume ellipsoid estimates of μ

and Σ proposed by Rousseeuw and van Zomeren (1990) or the estimates proposed by Hadi (1992, 1994)). In this case, the λ_j's are replaced by the eigenvalues of the estimated covariance matrix and u_j is replaced by the value on the jth principal component associated with an observation at **z**.

Example 4.4. Estimation of Binomial Parameters. Let $x_1, x_2, \ldots, x_k$ be k independent observations from a binomial random variable with parameters N and p. The problem of estimating N when p is also unknown has been considered by many authors; see, e.g., Olkin, Petkau, and Zidek (1981), Carroll and Lombard (1985), Casella (1986), and the references therein. It is well know that both the maximum likelihood and the method of moments estimators of N can be very highly unstable particularly in cases where N is large and p is small. For example, the method of moments estimators are given by

$$\hat{N} = \frac{\hat{\mu}^2}{\hat{\mu} - \hat{\sigma}^2}$$

and

$$\hat{p} = \frac{\hat{\mu}}{\hat{N}},$$

where

$$\hat{\mu} = k^{-1} \sum_{i=1}^{k} x_i \quad \text{and} \quad \hat{\sigma}^2 = k^{-1} \sum_{i=1}^{k} (x_i - \hat{\mu})^2.$$

Thus, if $\hat{\sigma}^2 > \hat{\mu}$, then $\hat{N} < 0$, which is unrealistic. If $\hat{\sigma}^2 < \hat{\mu}$, then $\hat{N}$ will be unstable when $\hat{\mu}$ is close to $\hat{\sigma}^2$. This case is likely to appear when p is small and N is large.

Suppose now that we wish to study the influence of an observation x_j on the stability of the estimate $\hat{N}$. We define

$$\hat{N}(\phi) = \frac{[\hat{\mu}(\phi)]^2}{\hat{\mu}(\phi) - \hat{\sigma}^2(\phi)},$$

where

$$\hat{\mu}(\phi) = (k - \phi)^{-1} \left[(1 - \phi) x_j + \sum_{i \neq j} x_i \right]$$

and

$$\hat{\sigma}^2(\phi) = (k - \phi)^{-1} \left[(1 - \phi)(x_j - \hat{\mu}(\phi))^2 + \sum_{i \neq j} (x_i - \hat{\mu}(\phi))^2 \right]$$

are the estimate of $\mu = Np$ and $\sigma^2 = Np\,(1 - p)$ when the weight $(1 - \phi)$ is assigned to observation x_j, respectively. With $\rho(\phi) = \hat{N}(\phi) - \hat{N}$, straightforward calculations yield

$$SF(M, T, P) = SF(x_j, \hat{N})$$

$$= \frac{2\hat{\mu}\, SF(x_j;\hat{\mu})(\hat{\mu} - \hat{\sigma}^2) - \hat{\mu}^2[SF(x_j;\hat{\mu}) - SF(x_j;\hat{\sigma}^2)]}{(\hat{\mu} - \hat{\sigma}^2)^2}$$

$$= \frac{(2\hat{p} - 1)\, SF(x_j;\hat{\mu}) - SF(x_j;\hat{\sigma}^2)}{\hat{p}^2},$$

where $SF(x_j;\hat{\mu}) = (\hat{\mu} - x_j)^2 / k$ and $SF(x_j;\hat{\sigma}^2) = [\hat{\sigma}^2 - (\hat{\mu} - x_j)^2] / k$ are the sensitivity functions for $\hat{\mu}$ and $\hat{\sigma}^2$, respectively. Note here that $SF(x_j;\hat{\mu})$ reflects the fact that an estimate of μ will increase when an observation less than $\hat{\mu}$ is downweighted, and decrease when an observation larger than $\hat{\mu}$ is downweighted. Similarly, an estimate of σ^2 will decrease if an observation far from the mean is downweighted, and increase if an observation close to the mean is downweighted. We also note that the effects on $\hat{\mu}$ and $\hat{\sigma}^2$ are reduced when the sample size k increases.

Example 4.5. Tests of Hypotheses. This example shows that an application of the sensitivity function, for assessing the effect of an additional parameter, yields the score (Lagrange multiplier) test. Suppose that a vector of random variables $\mathbf{Y}$ has probability density function $f(\mathbf{y}; \theta)$, where the $p \times 1$ parameter vector θ can be partitioned as $\theta = (\psi^T, \phi^T)^T$ and ϕ has dimension $q \le p$, so that the log likelihood function is $l(\psi, \phi) = \log f(\mathbf{y}; \theta)$. Suppose further that the information matrix exists and has an inverse partitioned as

$$\mathbf{i}_{\theta\theta}^{-1} = \begin{pmatrix} \mathbf{i}_{\psi\psi}^{-1} & \mathbf{i}_{\psi\phi}^{-1} \\ \mathbf{i}_{\phi\psi}^{-1} & \mathbf{i}_{\phi\phi}^{-1} \end{pmatrix}.$$

The model M_0 under consideration is obtained by restricting the parameter vector to $\theta = (\psi^T, \mathbf{0}^T)^T$. To assess the local sensitivity of the log likelihood when ϕ is perturbed around $\phi = 0$, we take $\rho(\phi) = P_T(M_\phi) - P_T(M_0)$, where $P_T(M_\phi) = \sup_{\psi} l(\psi, \phi) = l(\hat{\psi}_\phi, \phi)$ is the marginal likelihood under the model M_ϕ, ϕ being fixed, and $\hat{\psi}_\phi$ is the maximum likelihood estimate of ψ under M_ϕ. The definition of the sensitivity function and an application of the chain rule yield

$$\mathbf{SF} = \left[\frac{\partial\psi}{\partial\phi}\bigg|_{(\hat{\psi}_0, 0)}\right]^T \mathbf{U}_\psi(\hat{\psi}_0, 0) + \mathbf{U}_\phi(\hat{\psi}_0, 0),$$

where $\mathbf{U}_\psi = \partial l(\psi, \phi) / \partial \psi$ and $\mathbf{U}_\phi = \partial l(\psi, \phi) / \partial \phi$ are the efficient scores. By definition, $\mathbf{U}_\psi(\widehat{\psi}_0, 0) = \mathbf{0}$ so that $\mathbf{SF} = \mathbf{U}_\phi(\widehat{\psi}_0, 0)$. Since $\mathbf{SF}$ is a $q \times 1$ vector, it is useful to consider a semi-norm such as $\| \mathbf{SF} \|^2 = \mathbf{SF}^T \mathbf{M}\, \mathbf{SF}$, which is determined by a symmetric, positive (semi-) definite $q \times q$ matrix $\mathbf{M}$. A convenient norm in this case is obtained by letting $\mathbf{M} = \mathbf{i}^{-1}_{\phi\phi}$, thus yielding

$$Q = \mathbf{SF}^T \mathbf{i}^{-1}_{\phi\phi}\, \mathbf{SF},$$

which is recognized as the test statistic used in the score test (also known as the Lagrange multiplier test) for testing the hypothesis H_0: $\phi = \mathbf{0}$. We therefore conclude that S has an asymptotic χ^2 distribution with q degrees of freedom under H_0.

It is interesting to note here the differences among the score test, the Wald test, and the likelihood ratio test, all are testing the hypothesis H_0: $\phi = \mathbf{0}$. The three tests are illustrated graphically in Figure 3. Since $\mathbf{SF}$ is the derivative of $l(\widehat{\psi}_\phi, \phi)$ at $\phi = \mathbf{0}$, then the score test is the normed version of this derivative. On the other hand, the likelihood ratio test,

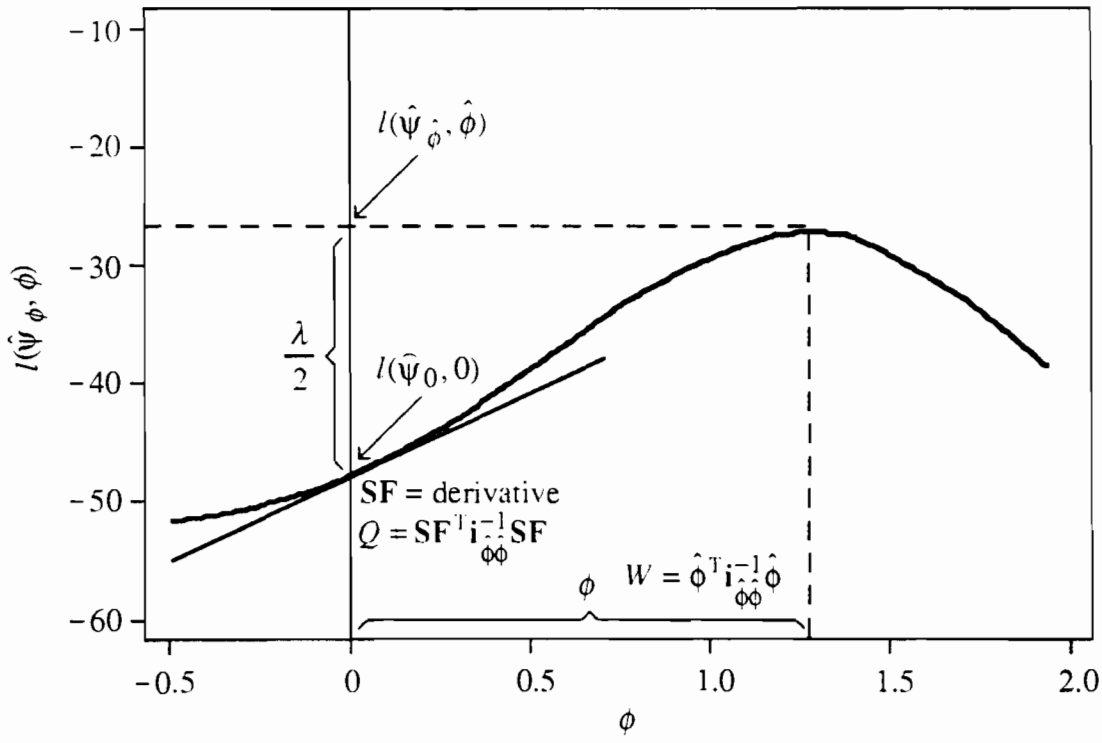

Figure 3. A graphical illustration of the core, the Wald, and the likelihood ratio tests. The score test, Q, is a normed version of $\mathbf{SF}$ which is the derivative of $l(\hat{\psi}_\phi, \phi)$ at $\phi = 0$. The Wald test, W, is a normed version of the MLE, $\hat{\phi}$, which is the horizontal distance between $\mathbf{0}$ and $\hat{\phi}$. The likelihood ratio test, λ, is twice the vertical distance between $l(\hat{\psi}_{\hat{\phi}}, \hat{\phi})$ and $l(\hat{\psi}_0, 0)$.

$$\lambda = 2\,(l(\hat{\psi}_{\hat{\phi}}, \hat{\phi}) - l(\hat{\psi}_0, \mathbf{0})),$$

is twice the difference (the vertical distance) between $l(\hat{\psi}_{\hat{\phi}}, \hat{\phi})$ and $l(\hat{\psi}_0, \mathbf{0})$, $\hat{\phi}$ being the maximum likelihood estimate of ϕ, whereas the Wald test,

$$W = \hat{\phi}^{\mathrm{T}} (\mathbf{i}^{-1}_{\phi\phi})^{-1} \hat{\phi},$$

is a normed version of the difference (the horizontal distance) between $\hat{\phi}$ and $\mathbf{0}$.

We should also note that, in cases where $\mathbf{i}_{\phi\phi}$ is singular, the tests can still be applied with an appropriate choice of a generalized inverse of $\mathbf{i}_{\phi\phi}$. See Hadi and Wells (1990, 1991) for details.

Example 4.6. Linear Regression. Consider the usual linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{4.7}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of a response variable, $\mathbf{X}$ is $n \times k$ matrix of predictors with rank $k < n$, β is a $k \times 1$ vector of parameters, and ε is an $n \times 1$ vector of random disturbances which are usually assumed to be independently and identically distributed as $N(0, \sigma^2)$.

Several statistics are usually of interest here, for example, the estimated parameter vector, $\hat{\beta}$, the vector of fitted values $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$, the fitted value at the point $\mathbf{x}_i$, $\hat{y}_i = \mathbf{x}_i^{\mathrm{T}}\hat{\beta}$, the residual mean square, $\hat{\sigma}^2$, etc. We wish to assess the influence of an observation $\mathbf{z}_i = (\mathbf{x}_i^{\mathrm{T}}, y_i)^{\mathrm{T}}$ on these statistics. Here we consider only the fitted value at the point $\mathbf{x}_i$, $\hat{y}_i = \mathbf{x}_i^{\mathrm{T}}\hat{\beta}$ because it leads to a simple diagnostic tool, but similar calculations can be performed for the other statistics.

We now examine the effects of $\mathbf{z}_i$ on the sampling distribution of $\hat{y}_i$. Let $P_T(M_\phi) = F(\phi)$ be the sampling distribution of $\hat{y}_i$ with a weight of $1 - \phi$ assigned to a perturbation at $\mathbf{z}_i$. As in Example 4.3, we use the Frechet distance to measure the difference between $F(\phi)$ and $F(0)$, the sampling distributions of $\hat{y}_i$ in the perturbed and in the unperturbed cases, respectively. In particular, this means that $F(1)$ is the sampling distribution of $\hat{y}_i$ when the observation $\mathbf{z}_i$ has been removed from the data set. Let $e_i = y_i - \mathbf{x}_i^{\mathrm{T}}\beta_0$, $p_{ii} = \mathbf{x}_i^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_i$, $p(\mathbf{x}_i, \phi) = \mathbf{x}_i^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} - \phi\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})^{-1}\mathbf{x}_i$, and $\sigma^2_\phi = (1 - \phi)\sigma^2_0 + \phi e_i^2$. After some algebra, one can show that

$$\mathbf{x}_i^{\mathrm{T}}\beta_\phi = \mathbf{x}_i^{\mathrm{T}}\beta_0) + \frac{\phi p_{ii} e_i}{1 - \phi(1 - p_{ii})} \tag{4.8}$$

and

$$p(\mathbf{x}_i, \phi) = \frac{p_{ii}}{1 - \phi p_{ii}}. \tag{4.9}$$

Then, under normality assumption,

$$F(\phi) = N(\mathbf{x}_i^{\mathrm{T}} \beta_\phi, \sigma_\phi^2 p(\mathbf{x}_i, \phi)). \tag{4.10}$$

Hence,

$$F(0) = N(\mathbf{x}_i^{\mathrm{T}} \beta_0, \sigma_0^2 p(\mathbf{x}_i, 0)). \tag{4.11}$$

These are univariate distributions. Thus, letting $G = F(\phi)$ and $H = F(0)$, (4.4) becomes

$$d(F(\phi), F(0)) = \left\{ \left[\mathbf{x}_i^{\mathrm{T}} (\beta_\phi - \beta_0) \right]^2 + \left[\sigma_\phi \sqrt{p(\mathbf{x}_i, \phi)} - \sigma_\phi \sqrt{p_{ii}} \right]^2 \right\}^{1/2}$$

$$= \left\{ \left[\frac{\phi p_{ii} e_i}{1 - \phi(1 - p_{ii})} \right]^2 + \left[\sqrt{\frac{p_{ii}\left[(1 - \phi)\sigma_0^2 + \phi e_i^2 \right]}{1 - \phi p_{ii}}} - \sqrt{p_{ii}\sigma_0^2} \right]^2 \right\}^{1/2}, \tag{4.12}$$

From (2.2) and (4.12) it follows that

$$SF(M, \hat{y}_i, P) = \lim_{\phi \to 0} \frac{1}{\phi} d(F(\phi), F(0))$$

$$= \left\{ p_{ii}^2 e_i^2 + \frac{p_{ii}\left[e_i^2 - \sigma_0^2 (1 - p_{ii}) \right]^2}{4\sigma_0^2} \right\}^{1/2}. \tag{4.13}$$

It should be noted here that the measure in (4.13) captures both the change in the fitted value itself as well as the change in its variance. An estimate of $SF(M, \hat{y}_i, P)$ in (4.13) can be obtained when the parameters are replaced by appropriate sample values.

Example 4.7. The Effect of an Additional Regressor. Suppose M_0 is a linear regression model as defined in (4.7). Suppose also that M_ϕ, is equal to M_0 except for (4.7) which is replaced by

$$\mathbf{Y} = \mathbf{X}\beta + \phi \mathbf{z} + \varepsilon, \tag{4.14}$$

where ϕ is an additional regression parameter and $\mathbf{z}$ is an $n \times 1$ vector containing the values of an additional regressor variable.

Let $\theta = (\beta^{\mathrm{T}}, \sigma^2, \phi)^{\mathrm{T}}$ and suppose the goodness-of-fit is measured by the log likelihood value, i. e., $P_7(M_\phi)$ is taken to be

$$\sup_{\beta, \sigma^2} l(\beta^{\mathrm{T}}, \sigma^2, \phi) = l(\hat{\beta}_\phi, \hat{\sigma}_\phi^2, \phi).$$

which is the largest value of the log likelihood function under the model M_ϕ, ϕ being fixed, and $\hat{\beta}_\phi$ and $\hat{\sigma}^2_\phi$ are the maximum likelihood estimates of β and σ^2 under M_ϕ. In this case, the sensitivity function can be written as

$$SF = \lim_{\phi \to 0} \{l(\hat{\beta}_\phi, \hat{\sigma}^2_\phi, \phi) - l(\hat{\beta}_0, \hat{\sigma}^2_0, \phi)\} / \phi.$$

An application of the chain rule yields

$$SF = \sum_{j=1}^{k} U_{\beta_j}(\hat{\beta}_0, \hat{\sigma}^2_0, 0) \frac{\partial \beta_j}{\partial \phi}\bigg|_{(\hat{\beta}_0, \hat{\sigma}^2_0, 0)} + U_{\sigma^2}(\hat{\beta}_0, \hat{\sigma}^2_0, 0) \frac{\partial \sigma^2}{\partial \phi}\bigg|_{(\hat{\beta}_0, \hat{\sigma}^2_0, 0)} + U_\phi(\hat{\beta}_0, \hat{\sigma}^2_0, 0),$$

where

$$U_{\beta_j}(\beta^T, \sigma^2, \phi) = \partial l(\beta^T, \sigma^2, \phi) / \partial \beta_j$$

$$U_{\sigma^2}(\beta^T, \sigma^2, \phi) = \partial l(\beta^T, \sigma^2, \phi) / \partial \sigma^2, \text{ and}$$

$$U_\phi(\beta^T, \sigma^2, \phi) = \partial l(\beta^T, \sigma^2, \phi) / \partial \phi$$

are the efficient scores. By definition, $U_{\beta_j}(\beta^T, \hat{\sigma}^2_0, 0) = U_{\sigma^2}(\beta^T, \hat{\sigma}^2_0, 0) = 0$, so that

$$SF = U_\phi(\hat{\beta}_0, \hat{\sigma}^2_0, 0) = \mathbf{z}^T\mathbf{e} / \hat{\sigma}^2_0,$$

where $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\beta}_0$ is the residual vector obtained from fitting model (4.7). By noting that $n\hat{\sigma}^2_0 = \mathbf{e}^T\mathbf{e}$ it follows that $SF = n\,(\mathbf{e}^T\mathbf{e})^{-1}\mathbf{e}^T\mathbf{z}$, which is recognized as the estimated regression coefficient when the additional variable $\mathbf{z}$ is regressed on the residuals from the model M_0.

For this model we have $i_{\phi\phi} = (\mathbf{z}^T\mathbf{z} - \mathbf{z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z})/\sigma^2_0 = \mathbf{u}^T\mathbf{u}/\sigma^2_0$, where $\mathbf{u}$ is the residual vector obtained when $\mathbf{z}$ is regressed on $\mathbf{X}$. Thus, estimating σ^2_0 by $\hat{\sigma}^2_0$, an estimate of the proposed norm of SF is

$$V = (SF)^2\, i^{-1}_{\phi\phi} = \frac{n\hat{\phi}^2\mathbf{e}^T\mathbf{e}}{\mathbf{u}^T\mathbf{u}}, \tag{4.15}$$

which is recognized as the score test for testing the hypothesis H_0: $\phi = 0$. We, therefore, conclude that V has an asymptotic χ^2 distribution with one degree of freedom under H_0.

A special case of (4.15) is the mean-shift outlier model, in which $\mathbf{z}$ is the ith unit vector (i. e., the ith observation has a different intercept from the remaining observations). For this model we find that $SF = e_i / \hat{\sigma}^2_0$ and $i_{\phi\phi}=(1 - p_{ii}) / \hat{\sigma}^2_0$, where p_{ii} is the ith diagonal element in the projection matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Therefore, V in (4.15) reduces to $r_i^2 = e_i^2 / \hat{\sigma}^2_0\,(1 - p_{ii})$, which is the squared internally studentized residual. Thus

diagnostics based on residuals point out that either the observation is an outlier or the null model is unsuitable. or both. In other words, diagnostics based on the residuals alone may identify wrong observations rather than wrong models, see also Schwarzmann (1991). Therefore, only further investigations and more knowledge about the subject matter can indicate the proper alternative.

5. Summary and Conclusions.

Since statistical models are simplification of reality, they are almost always wrong. A careful statistical analysis must take this fact into account. Therefore , statistical analysis should be viewed as an iterative process; a process which takes inputs and produces outputs in an iterative or cyclical way. Before reaching a final conclusions, the outputs at a given cycle is used to criticise and possibly ratify the inputs. This iterative process is discussed in section 2 and schematically shown in Figure 2. The sensitivity function is introduced in Section 3 as a general framework for assessing the sensitivity of the outputs to small changes in the input at a given cycle in the statistical process. Several examples from various areas in statistics illustrating the general applicability of the sensitivity function is given in Section 4. Some of these applications lead to known statistical techniques, for example, the influence function and the score test; while other applications can produce new ones, e.g., (4.6) and (4.13).

References

Atkinson, A. C. (1985), *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford: Clarendon Press.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Multicollinearity*, New York: John Wiley & Sons.

Box, G. E. P. (1979), "Robustness in the Strategy of Scientific Model Building," in *Robustness in Statistics*, R. L. Launer and G. N. Wilkinson, eds., New York: John Wiley & Sons.

Box, G. E. P. (1979), "Sampling and Bayes' Inference in Scientific Modeling and Robustness," *Journal of the Royal Statistical Society*, Series (A), 143, 383-430.

Carroll, R. J. and Lombard, F. (1985), "A Note on N Estimators for the Binomial Distribution ," *Journal of the American Statistical Association*, 80, 423-426.

Casella, G. (1986), "Stabilizing Binomial n Estimators," *Journal of the American Statistical Association*, 81, 172-175.

Chatterjee, S. and Hadi, A. S. (1988), *Sensitivity Analysis in Linear Regression*, New York: John Wiley & Sons.

Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15-18.

Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman and Hall

Davison, A. C. and Tsai, C.-L. (1992), "Regression Model Diagnostics," *International Statistical Review*, 60, 3, 337-353.

Dowson, D. C. and Landau, B. V. (1982), "The Frechet Distance Between Multivariate Normal Distributions," *Journal of Multivariate Analysis*, 12, 450-455.

Frechet, M. (1957), "Sur la Distance de Deux Lois de Probabilite," *C. R. Acad. Sci. Paris*, 244, 689-692.

Hadi, A. S. (1992), "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society, Series (B)*, Vol. 54, No. 3, 761-771.

Hadi, A. S. (1994), "A Modification of a Method for the Detection of Outliers in Multivariate Samples," *Journal of the Royal Statistical Society, Series (B)*, 56, 393-396.

Hadi, A. S. and Wells, M. T. (1990), "A Note on Generalized Wald's Test," *Metrika*, 37, 309-315.

Hadi, A. S. and Wells, M. T. (1991), "Minimum Distance Method of Estimation and Testing When Statistics Have Limiting Singular Multivariate Normal Distribution," *Sankhyā*, Series B, Part 2, 53, 257-267.

Hampel, F. R. (1968), "Contributions to the Theory of Robust Estimation," Ph.D. thesis, University of California, Berkeley.

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383-393.

Hampel, F. R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley Sons.

Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley & Sons.

Nyquist, H. (1992), "Sensitivity Analysis in Empirical Studies," *Journal of Official Statistics*, 8, 167-182.

Olkin, I., Petkau, A. J., and Zidek, J. V. (1981), "A Comparison of n Estimators for the Binomial Distribution," *Journal of the American Statistical Association*, 76, 375-642.

Pregibon, D. (1981), "Logistic Regression Diagnostics," *The Annals of Statistics*, 9, 705-724.

Schwarzmann, B. (1991), "A Connection Between Local-Influence Analysis and Residual Diagnostics," *Technometrics*, 33, 1, 103-104.